Adrar University
Faculty of Science and Technology
Department of Mathematics and Computer Science

Initiation to Research (Course)

2nd Year Master (S3)
2020/2021

Build a parallel corpus (Algerian Dialect, Modern Standard Arabic) and apply Deep Neural Machine Translation on it

Elwannas HIRI [1]

Instructor: Dr. Abdelghani DAHOU [2]

February 15, 2021

[1] Email: hiri.elwannas@gmail.com
[2] Email: dahou.abdghani@univ-adrar.edu.dz

CONTENTS

LIST OF FIGURES

1 ABSTRACT

Arabic Dialects are getting under search radar lately, especially in the Artificial Intelligent (AI) field, the last one requires a huge amount of data to learn from it. Modern Standard Arabic used in official newspapers, Conferences, books and more. On the other, hand we have various dialects. Some of them are similar to the Moroccan dialect(MD) and some of them are similar to Tunisian dialects(TD) and the others are just Algerian original dialects. In this research proposal, we first present how we built a 13,000 parallel Algerian Dialects(AD) to Modern Standard Arabic(MSA) sentences. Next, we present our first experiment using FAIRSEQ(7) toolkit on our data set, and why we used FAISEQ(7) instead of others(8). The first experimental results was 1.02 using the BLUE4 Score without any modification on the data set. Finally a short description of future work to get a better result.

2 INTRODUCTION

Sentences for each file, we had created a help post in Facebook help community, also posted in 1001Tech[3] group, received 126+ files from all the Algeria regions. To boost the process we paid 4 people to write 6400 sentences. After collecting the data and organize it we got 14,000 sentences (until 14-02-2021). Since 2010, several works are using Statistical Machine Translation(SMT), SRILM, and Other approaches such as machine translation system that uses an MSA pivot approach(1). Lately, deep neural networks came to the machine translation field. Many approaches have been developed such as Deep Neural Network(DNN), Feed Forward Neural Network(FFNN), Conventional Neural-Network(CNN), Recursive Neural Network (RNN) and finally Purely Neural Network (8). Available DNN toolkits (e.g. Open-NMT[4], FAIRSEQ(7)) are widely used by big companies(e.g. Google Translate, Facebook). The study will help to increase the interaction between Arab countries. In recent years topics like Arabic Dialects are widely discussed in social media, this results increase the attention between Arabs them selves. The main motivation for the creation of parallel corpus is Dial2MSA(2), on the other hand the power of current machine translation is a good motivation for the Machine Translation side. Previous works (1) (4) (3) used up in years methods. We took the state of the art which is FAIRSEQ(7), we applied Transformer model on our data set, we got a 1.02 BLUE4 score for the first experiment. For the future work, we are going to extend the data using some methods, normalize data and try to get better results.

3 RELATED WORKS

3.1 PADIC(1)

PADIC(1) is a 6,400 sentences corpus of five different dialects Annaba(Algerian city), Algiers(Algeria Capital), Palestinian, Tunisian and Syrian Dialects with MSA parallel sentences. Algerian Dialect is non-resourced language. The PADIC corpus was built from scratch. Using

[3]Facebook Group: https://www.facebook.com/groups/dztech1001/
[4]Open-NMT: A modern toolkit for machine translation

Kneser-Ney and Witten-Bell smoothing techniques the BLUE results gotten from translating AD to MSA were respectively 15.1, 14.64. Sentences was written in Transliteration(1).

3.2 DIAL2MSA (2)

Dial2MSA(2) was the first multi dialects corpus. The Corpus contains 5500 tweets written in Egyptian, Gulf, Levantine, Maghribi with MSA translations. Twitter was the source of data, 175M tweets have been filtered with powerful words of the dialects. CrowedFlower[5] was the tool controlling the quality of data annotation(2).

3.3 "A TRANSLATOR FOR ARABIC DIALECTS TO MODERN STANDARD ARABIC" (3)

The Study (3) was about applying Morpho-Semantic analysis, Morphological analyser to obtain the translation.

3.4 "A HYBRID APPROACH TO TRANSLATE MOROCCAN ARABIC DIALECT" (4)

Arabic Dialects are characterized by diglossia[6]. MGB Dialect uses words from other langauge for example Spanich, French, Amazigh and MSA. By applying the morphological analysis and rule-based translation system with smoothing using statistical tool the performance of translation get better. Sentences were written in transliteration form using Buckwalter.

3.5 "AN ALGERIAN DIALECT STUDY AND RESOURCES" (5)

This work is a part of TORJOMAN(9), an analysis for Algiers Dialect. The work present Phonological differences, types of verbs, negation form in the Algerian day to day speech, types of particles and Inflection of Algerian Dialect from MSA (5).

3.6 "PARALLEL RESOURCES FOR TUNISIAN ARABIC DIALECT TRANSLATION" (6)

A combined work of three corpora of dialects to translate Tunisian Dialect to MSA. The Research was focused on getting the largest possible TD-MSA corpus to apply Machine Translation on the data.Three experiments was tried, the first one got 12 on BLUE score, the result was caused by the difference between vocabulary used in corpora, the last one reach 15.03 BLUE score after the adjustment.

[5] CrowedFlower website has been changed to: https://visit.figure-eight.com/

[6] diglossia: A situation in which two languages (or two varieties of the same language) are used under different conditions within a community, often by the same speakers.

4 METHODOLOGY/RESEARCH METHODS

The workflow of this research is presented below:

- Build corpus
  - Split the data set
  - Distribute it the largest possible regions in Algeria to get more data
- Normalizing the data set
- Create a vocabulary using subword-nmt[7]
- Train the data using FAIRSEQ(7) with Transformer model.
- Adjust parameters for better results.

Previous presented related works used old methods for Machine Translation. We are entering the AI field on this research. Lately Deep Neural Network entred all research fields like Image processing Speech Recognition...etc. In 2017, Google Translate has used the OpenNMT. All Methods of Deep Learning Neural Network Machine Translation are presented in (8).

4.1 BUILDING THE CORPUS

Using the free Data set from Tatoeba, we split it into 100 sentences for each excel file to make the translations easier. More than 150 person helped to build the corpus from 22 different cities. Adrar and Tipaza were the most active cities in this translations process. 126 files has been received. .txt,.csv,.xlxs are files types received. everyone can use the corpus after publishing the article.

4.2 NORMALIZING THE DATA

The process of Normalizing the data is simple for the first experiment just removing the numbers, dots, commas and repeated sentences.

4.3 CREATE VOCABULARY USING SUBWORD-NMT

Subword-nmt is the main state of the are used techniques to create the vocabulary before training the data in Deep Neural Network Machine Translation systems. By fixing the vocab to 2000 words, the vocabulary is ready to be used by the model.

4.4 TRAIN THE DATA SET USING FAIRSEQ(7)

FAIRSEQ(7) is newest toolkit developed by Facebook under Pytorch library. we choose FAIRSEQ(7) because of it performance and large support of models in addition tothe ability to online back translation and EM-style(10) training to produce lower the loss.

[7] subword-nmt: Byte Pair Encoding method to create a vocabulary in order to train the data

4.5 RESULTS

For the model Transformer, we used these parameters:

- 3 by 3 for the hidden layers
- batch 400
- learning rate 1e-3

Caused by the amount of resources

- Drop out 0.01
- Adam for smoothing
- Token size 200

The result we got is 1.02 using BLUE4 evaluation and it is explained by:

- Lack of DA resources

Neural networks requires a big amount of data to get better results

- Different Dialects
- Simple Normalization

5 FUTURE WORK

Our Future work is based on:

- extending the corpus by adding PADIC(1),Dial2MSA (2) , Tunisian Dialect corpus and Moroccan resources.
- Normalizing the data using different available methods.
- Try different models than Transformer to our data set.
- Try models on our data set based on regions.

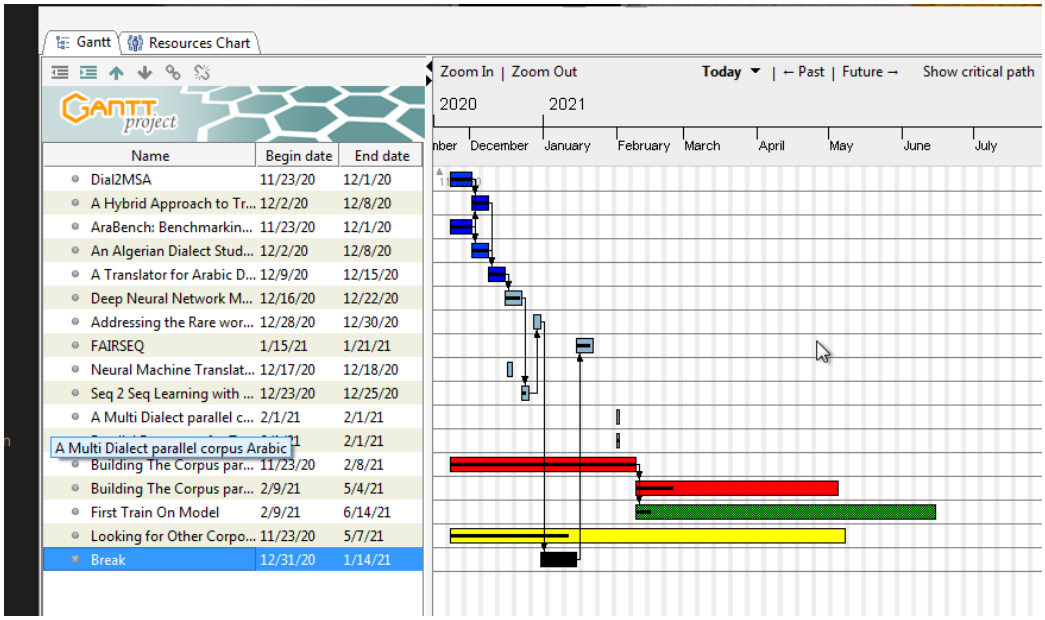

Figure 6.1: WorkFlow Time Line

6 Project Timeline

This work has Started from November 22th, 2020 it started by exploiting current researches and techniques. For now it is all about learning availables models , trying normalization methods and adjusting the parameters and getting better results. Building the corpus takes forever.

REFERENCES

1 MEFTOUH, K. et al. Machine translation experiments on padic: A parallel arabic dialect corpus. In: *The 29th Pacific Asia conference on language, information and computation*. [S.l.: s.n.], 2015.

2 MUBARAK, H. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), OSACT2018 Workshop*. [S.l.: s.n.], 2018. p. 49–53.

3 MAHGOUBA, H. E.; SHAABANB, Y. A translator for arabic dialects to modern standard arabic. In: *The International Workshop on Computers and Information Sciences (WCIS), At Tabuk, Kingdom of Saudi Arabia*. [S.l.: s.n.], 2015.

4 TACHICART, R.; BOUZOUBAA, K. A hybrid approach to translate moroccan arabic dialect. In: IEEE. *2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14)*. [S.l.], 2014. p. 1–5.

5 HARRAT, S. et al. An algerian dialect: Study and resources. *International journal of advanced computer science and applications (IJACSA)*, v. 7, n. 3, p. 384–396, 2016.

6 KCHAOU, S.; BOUJELBANE, R.; BELGUITH, L. H. Parallel resources for tunisian arabic dialect translation. In: *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. [S.l.: s.n.], 2020. p. 200–206.

7 OTT, M. et al. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

8 ZHANG, J.; ZONG, C. et al. Deep neural networks in machine translation: An overview. *IEEE Intell. Syst.*, v. 30, n. 5, p. 16–25, 2015.

9 HARRAT, S. et al. Building resources for algerian arabic dialects. In: *15th Annual Conference of the International Communication Association Interspeech*. [S.l.: s.n.], 2014.

10 SHEN, T. et al. *Style Transfer from Non-Parallel Text by Cross-Alignment*. 2017.